مجله دانشکده پزشکی، دانشگاه علوم پزشکی تهران، دوره ۶۴ ، شماره ۳، خرداد ۱۳۸۵: ۱۱-۵

بررسی پایایی رادیولوژیست‌ها وعملکرد آنها در تشخیص وخامت توده‌های تخمدان از روی سونوگرافی

دکتر علیرضا اکبرزاده باغبان[۱] *
دکتر غلامرضا بابایی[۲]
دکتر انوشیروان کاظم نژاد[۲]
دکتر سقراط فقیه زاده[۲]
دکتر فاطمه برادران انارکی[۳]
دکتر زهرا الهی پناه[۳]

۱- گروه آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی
۲- گروه آمار زیستی، دانشگاه تربیت مدرس
۳- گروه رادیولوژی، دانشگاه علوم پزشکی تهران

چکیده

زمینه و هدف: پایایی نظرات ارایه شده از مشاهده سونوگرافی مربوط به توده‌های تخمدان، دارای اهمیت زیادی می‌باشد. قابلیت تشخیص که یکی از موضوعات مهم مرتبط با این مبحث است، توانایی متخصصان در تشخیص صحیح وخامت توده‌ها را مورد ارزیابی قرار می‌دهد. در این مقوله می‌توان بررسی نمود که وضعیت وخامت، تا چه میزان برای رادیولوژیست‌ها قابل تشخیص می‌باشند.

روش بررسی: در تحقیق حاضر که از نوع مقطعی - تحلیلی می‌باشد، ۵ رادیولوژیست، شامل ۲ متخصص و ۳ دستیار به طور مستقل و در دو زمان مجزا (به فاصله یک هفته) وخامت توده تخمدان را برای ۴۰ سونوگرافی مشخص نمودند. این سونوگرافی‌ها در طول دی‌ماه سال ۱۳۸۳ از زنان مراجعه کننده به بخش رادیولوژی بیمارستان تخصصی زنان میرزاکوچک‌خان و به کمک یک دستگاه و یک رادیولوژیست گرفته شدند.

در این مقاله ابتدا به کمک ضریب کاپای موزون، پایایی اندازه‌گیری رادیولوژیست‌ها که معرف توافق داخلی آنها (توافق هر رادیولوژیست با خودش) می‌باشد، بررسی شده، سپس به کمک مدل پیوند مربع امتیازات و مدل توافق به‌علاوه پیوند مربع امتیازات، قابلیت تشخیص آنها در تعیین وخامت توده تخمدان مورد ارزیابی قرار گرفته است.

یافته‌ها: برای داده‌های مربوط به ۲ رادیولوژیست با عملکرد خوب، مدل پیوند مربع امتیازات برازش داشت. میانگین ضریب کاپای موزون برای این دو نفر برابر ۰/۸۱ و میانگین قابلیت تشخیص برای آنها عبارت بود از ۰/۹۹ . برای ۳ رادیولوژیست دیگر که عملکرد ضعیف‌تری داشتند، مدل توافق به‌علاوه پیوند مربع امتیازات مناسب بود. این دستیاران دارای میانگین ضریب کاپاموزون ۰/۶۵ و میانگین قابلیت تشخیص ۰/۹۷ بودند.

نتیجه‌گیری: اگرچه متخصصان رادیولوژی عملکرد بهتری نسبت به دستیاران رادیولوژی داشتند، همه آنها در تشخیص وضعیت وخامت توده تخمدان از روی سونوگرافی دارای کارکرد مناسبی بودند. به‌علاوه تفکیک سطوح خوش‌خیم از بینابین برای رادیولوژیست‌ها مشکل‌تر از تفکیک سطوح بدخیم و بینابین بود، اگر چه در این مورد نیز متخصصان بهتر از دستیاران بودند.

کلمات کلیدی: توده تخمدان، سونوگرافی، کاپای‌موزون، قابلیت تشخیص، مدل پیوند، پایایی

* ***نشانی:*** تهران، میدان قدس (تجریش)، ابتدای خیابان دربند، دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، گروه آمار زیستی، صندوق پستی ۴۶۱۸ - ۱۹۳۹۵، تلفن :۲۲۷۰۷۳۴۷، نمابر: ۲۲۷۲۱۱۵۰
پست الکترونیک: akbarzad@sbmu.ac.ir

مقدمه

فرض کنیم آزمایندهای (مثل یک رادیولوژیست) در دو زمان متفاوت هر کدام از افراد یک نمونه را طبق یک مقیاس ترتیبی (مثل وضعیت وخامت توده تخمدان) طبقهبندی میکند، به طوری که نتیجه طبقهبندی در زمان اول روی نتیجه زمان دوم هیچ تأثیری نداشته باشد. میتوان نتیجه توأم این دو بار تشخیص را در یک جدول متقاطع آورد و دو موضوع مهم را بررسی نمود : ۱- قابلیت تشخیص ردهها. زمانی که مقیاس دادهها ترتیبی است، قابلیت تشخیص ردهها یکی از موضوعات مهم قابل ارزیابی میباشد. این مقوله تعیین میکند که آیا ردههای مقیاس رتبهای برای آزماینده قابل تفکیک هستند یا خیر؟ [۱]. ۲- توافق نرخگذاریهای آزماینده در دو زمان اندازهگیری. این موضوع در واقع فراوانی روی قطر اصلی این جدول بوده و بیانگر پایایی اندازههای مربوط به هر آزماینده میباشد [۴-۲]. در این مقاله موضوع اول با استفاده از مدلهای آماری و موضوع دوم با استفاده از کاپایموزون بررسی میشوند.

اکثر مقیاسهای ترتیبی، ذهنی بوده و به عنوان مثال تشخیص این که وضعیت بیماری فردی در چه سطحی از یک مقیاس ترتیبی قرار دارد، غالباً مشکل بوده و حتی آزمایندههای خیلی مجرب، در دو بار اندازهگیری روی افراد مشترک نظرات متفاوتی ارایه میکنند [۳].

معمولاً برای بررسی پایایی، یعنی ارزیابی میزان توافق هر آزماینده با خودش[1] از ضرایب توافق کاپا یا کاپاموزون استفاده میشود [۶-۵]. استفاده از این ضرایب و تکیه صرف به نتایج حاصل از آنها، خالی از اشکال نمی باشد و محققین زیادی توصیه کردهاند که در کنار محاسبه آنها، ساختار توافق دادههای حاصل نیز مدلبندی شود و بررسی کاملتری انجام شود [۱، ۷، ۸] لذا در این مقاله در کنار محاسبه ضریب کاپایموزون، ساختار مذکور در ارتباط با شاخص ترتیبی وضعیت وخامت توده تخمدان [با سطوح خوشخیم، بین خوشخیم و بدخیم(بینابین)، بدخیم] مدلبندی گردیده و به کمک آن قابلیت تشخیص رادیولوژیستها درتفکیک سطوح این شاخص بررسی شده است.

روش بررسی

مطالعه انجام شده در این مقاله از نوع مقطعی-تحلیلی[2] میباشد که درآن برای ارزیابی قابلیت تشخیص رادیولوژیستها در تفکیک سطوح شاخص ترتیبی وضعیت وخامت توده تخمدان، از دو مدل آماری پیوند مربع امتیازات و توافق بهعلاوه پیوند مربع امتیازات استفاده شده است. مدلهای مذکور با اندک تغییری در امتیازات انتسابی به گروههای ترتیبی، به ترتیب ازمدل پیوند یکنواخت [۹] و مدل توافق بهعلاوه پیوند یکنواخت [۱۰، ۱۱] به دست میآیند.

دادههای این مقاله از بخش رادیولوژی بیمارستان تخصصی زنان میرزاکوچکخان جمعآوری شدند. بعد از اخذ رضایت ۴۰ زن مراجعه کننده، علاوه برتحویل سونوگرافی تخمدان بیمار به او، نسخهٔ دیگری نیز برای انجام تحقیق گرفته میشد. همه این سونوگرافیها توسط یک دستگاه و یک رادیولوژیست گرفته شد تا تغییر شرایط روی نتیجه نهایی تأثیر منفی نداشته باشد. این سونوگرافیها به طور جداگانه به هر کدام از ۵ رادیولوژیست (شامل ۲ متخصص و ۳ دستیار) داده شد تا تشخیص خود را با اختصاص کدهای یک تا سه، به ترتیب برای خوشخیم تا بدخیم گزارش نمایند. بعد از گذشت یک هفته و به صورت یک سوکور، این سونوگرافیها مجدداً به همان رادیولوژیستها داده شد تا دوباره آنها را کدگذاری

1 - Intra-rater agreement

2- Analytic Cross-Sectional

جدول ۱ - ضریب کاپای موزون و قابلیت تشخیص برای ۵ جدول ۳×۳ حاصل از نرخ گذاری هرکدام از رادیولوژیست‌ها در دو زمان مختلف

شماره رادیولوژیست	کاپای‌موزون	قابلیت تشخیص برای سطوح بدخیم و بینابین	قابلیت تشخیص برای سطوح بینابین و خوش‌خیم
۱*	۰/۶۱	۰/۹۴۲۳	۰/۹۷۱۷
۲*	۰/۶۹	۰/۹۶۰۱	۰/۹۸۶۴
۳*	۰/۶۵	۰/۹۶۴۰	۰/۹۷۶۴
۴**	۰/۷۵	۰/۹۸۶۷	۰/۹۹۹۹
۵**	۰/۸۷	۰/۹۹۴۰	۰/۹۹۹۹

* مدل توافق به‌علاوه پیوند مربع امتیازات (مربوط به دستیاران)

** مدل پیوند مربع امتیازات (مربوط به متخصصان)

نمایند. فرصت یک هفته‌ای از طرفی برای از بین رفتن ذهنیت قبلی رادیولوژیست‌ها کافی به نظر می‌رسد و از طرف دیگر آن‌قدر طولانی نیست که روی کیفیت سونوگرافی‌ها تأثیر منفی بگذارد. نرخ‌گذاری متقاطع هر کدام از این رادیولوژیست‌ها در دو زمان، منجر به تولید ۵ جدول ۳×۳ شد که این جداول مبنای عمل این مقاله قرار گرفتند.

برای انجام تحقیق روی داده‌های حاصل، بعد از محاسبه ضریب کاپای‌موزون، برازندگی[1] مدل‌های مورد نظر بررسی شد. سپس بهترین مدل با استفاده از آماره نیکویی برازش[2] انتخاب شد و از روی آن قابلیت تشخیص هرکدام ازرادیولوژیست‌ها محاسبه گردید.

در این تحقیق ۴۰ سونوگرافی در دو زمان توسط هر کدام از رادیولوژیست‌ها مورد ارزیابی قرار گرفتند. با توجه به این که حجم نمونه در مطالعات مربوط به ارزیابی روایی و پایایی نوعاً کم است (۱۵ الی ۲۰ نمونه برای متغیرهای کمی و اندکی بیشتر برای متغیرهای کیفی)، این تعداد می‌تواند نمونه قابل قبولی بوده و زیر بنای انجام نتیجه‌گیری‌های آماری قرار گیرد [۱۲].

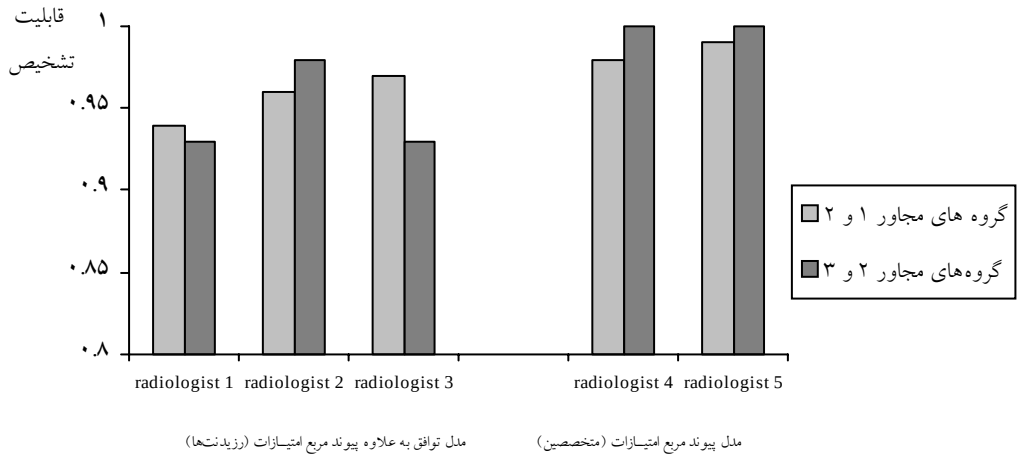
قابلیت تشخیص برای هر رادیولوژیست در ارتباط با دو سطح مجاور شاخص وضعیت وخامت توده تخمدان بیان می‌دارد که آن دو سطح تا چه میزان برای آن رادیولوژیست قابل تفکیک می‌باشند [۱۳، ۱۴]. دامنه تغییرات این شاخص مانند ضریب تعیین (R^2) در مدل‌های رگرسیون بین صفر و یک می‌باشد، بنابراین هر قدر قابلیت تشخیص دو رده از این شاخص به عدد ۱ نزدیک‌تر باشد، می‌گوییم آن رادیولوژیست در تفکیک دو رده مذکور بهترعمل کرده است.

برای ورود داده‌ها و استخراج جداول دو بعدی از نرم‌افزار SPSS و برای محاسبه ضریب کاپای‌موزون، برازش مدل‌های مورد نظر روی داده‌های جداول مذکور و برآورد شاخص‌ها از نرم افزار SAS استفاده شد. محاسبه قابلیت تشخیص با استفاده از فرمول‌نویسی در نرم‌افزارEXCEL انجام شد و نمودار نیز با همین نرم‌افزار ترسیم گردید.

یافته‌ها

تحلیل این مقاله با شاخص ۳ حالته وضعیت وخامت توده تخمدان در قالب ۵ جدول انجام شد. مدل توافق به‌علاوه پیوند

1- Fitness
2- Goodness of Fit



نمودار ۱- قابلیت تشخیص رادیولوژیست‌ها به تفکیک مدل و گروه‌های مجاور

مربع امتیازات در ۳ مورد (مربوط به دستیاران) و مدل پیوند مربع امتیازات در۲ مورد (مربوط به متخصصان) بهترین برازش را نشان دادند. جدول ۱ خلاصه‌ای از کارهای انجام شده روی داده‌ها را نشان می‌دهد.

اطلاعات موجود در جدول ۱ نشان می‌دهد رادیولوژیست‌هایی که مدل پیوند مربع امتیازات برای نرخ‌گذاری آنها مناسب بود، دارای قابلیت تشخیصی بالا، با حداقل مقدار ۰/۹۸ برای گروه‌های مجاور ۱ و ۲ و حداقل مقدار ۰/۹۹ برای گروه‌های مجاور ۲ و ۳ بودند. برای متخصصان رادیولوژی، تفاوت قابل ملاحظه‌ای در قابلیت تشخیص گروه‌های ۱ و۲ با ۲ و ۳ وجود نداشت، یعنی آنها در تفکیک گروه‌های ۱ و ۲ همان قدر مهارت داشتند که در تفکیک گروه‌های ۲ و ۳، زیرا میانگین قابلیت تشخیص آنها برای گروه‌های ۱ و ۲ و همچنین برای گروه‌های ۲ و ۳ برابر۰/۹۹ به دست آمد.

قابلیت تشخیص برای سه رادیولوژیست دیگر(دستیاران)، که مدل توافق به‌علاوه پیوند مربع امتیازات برای نرخ‌گذاری آنها مناسب تشخیص داده شد، نسبتاً پایین‌تر بود. به ویژه برای گروه‌های مجاور ۱ و ۲، آنها دارای میانگین قابلیت تشخیص ۰/۹۵ بودند. این دستیاران در تفکیک سطوح ۲ و۳، با حد اقل قابلیت تشخیص ۰/۹۷ و میانگین قابلیت تشخیص ۰/۹۸ وضعیت بهتری نسبت به سطوح ۱ و ۲ داشتند. در مجموع دستیاران رادیولوژی با میانگین کلی ۰/۹۷ دارای قابلیت تشخیص پایین‌تری نسبت به متخصصان رادیولوژی، با میانگین کلی ۰/۹۹ بودند. نمودار ۱ جزییات بیشتری را ارایه می‌کند.

بحث

برای مقایسه کامل اندازه‌گیری‌های شاخص وضعیت وخامت توده تخمدان توسط رادیولوژیست‌ها در دو زمان و بررسی پایایی آنها، ابتدا کاپای‌موزون محاسبه گردید. همان‌گونه که از جدول ۱ ملاحظه می‌شود، اگر چه برای هیچ کدام از رادیولوژیست‌ها توافق تقریباً کاملی (به عنوان مثال با ضریب توافق حد اقل ۰/۹) بین نرخ‌گذاری‌ها در دو زمان دیده نمی‌شود، ولی این ضریب با تغییر از ۰/۶۱ (برای رادیولوژیست

شماره ۱) تا ۰/۸۶ (برای رادیولوژیست شماره ۵) و میانگین ۰/۷۱ مقادیر خوبی از توافق را ارایه می‌کند و مبین پایایی قابل قبول می‌باشد [۱۵]. یافته‌های این مقاله با نتایج حاصل از تحقیق آمر و همکاران [۱۶] که در آن میانگین توافق داخلی آزماینده‌ها را ۶۹/۴ درصد گزارش کرده‌اند و همچنین با نتایج حاصل از تحقیق استوکر و همکاران که در آن قدرت تشخیص سونوگرافی ۹۶/۲ درصد بوده است، هماهنگی دارد [۱۷].

با وجود این که دستیاران رادیولوژی دارای قابلیت تشخیص پایین تری نسبت به متخصصان رادیولوژی بودند، ولی این تفاوت خیلی چشمگیر نبود، زیرا همه آنها در تفکیک کلیه سطوح مجاور دارای قابلیت تشخیص حد اقل ۰/۹ بودند. اما در هر ۵ مورد قابلیت تشخیص برای گروه‌های مجاور ۱ و ۲ پایین‌تر از گروه‌های مجاور ۲ و ۳ به دست آمد. لذا به نظر می‌رسد که تفکیک سطوح مجاور پایین شاخص مذکور (خوش‌خیم و بینابین) برای رادیولوژیست‌ها مشکل‌تر از تفکیک سطوح مجاور بالای این شاخص (بینابین و بدخیم) باشد. اگرچه در این مورد نیز عملکرد متخصصان بهتر از دستیاران بود.

معمولاً برای بررسی روایی و پایایی اندازه‌گیری وضعیت وخامت توده تخمدان، از کاپا یا کاپای‌موزون استفاده می‌شود [۱۶]. اگرچه این ضرایب میزان کلی توافق را نشان می‌دهند، اما با توجه به اشکالاتی که در مقالات مختلف به آنها وارد شده [۸-۴] و نظر به این که نمی‌توان به کمک آنها در مورد قابلیت تشخیص آزماینده‌ها اظهارنظر نمود، استفاده از این ضرایب چندان خالی از ایراد نبوده و لازم است که با استفاده از مدل‌های آماری به مقوله قابلیت آزماینده‌ها در تفکیک رده‌های مقیاس ترتیبی نیز پرداخته شود تا بتوان در مراحل بعدی آموزش آزماینده‌ها ازاین یافته‌ها بهره برد.

در بررسی قابلیت تشخیص رده‌ها توسط رادیولوژیست‌ها در یک مقیاس ترتیبی، ابتدا ضرایب توافق کاپاموزون را برای جدول متقاطع حاصل از نرخ‌گذاری آنها در دو زمان محاسبه می‌نماییم. در صورت وجود توافق تقریباً کامل (مثلاً کاپاموزون بیشتر از ۰/۹) و قابل قبول بودن پایایی محاسبه شده، بدون انجام مدل‌بندی اظهار می‌داریم که رده‌ها به خوبی توسط رادیولوژیست‌ها قابل تشخیص می‌باشند. در غیر این صورت برای رسیدن به قابلیت تشخیص رده‌ها، می‌توان مدل‌های پیوند مربع امتیازات و توافق به‌علاوه پیوند مربع امتیازات را مورد نظر قرار داد. بعد از رسیدن به بهترین مدل و برآورد پارامترهای آن، قابلیت تشخیص رده‌ها به آسانی قابل برآورد و تفسیرخواهد بود.

چنانچه مدل پیوند مربع امتیازات به داده‌های رادیولوژیستی برازش داشته باشد، اظهار می‌داریم که آن رادیولوژیست دارای قابلیت تشخیص بالایی در تفکیک همه سطوح مجاور هم می‌باشد. در غیر این صورت بیان می‌کنیم که آزماینده در تشخیص سطوح مقیاس ترتیبی، به ویژه سطوح مجاور هم ابتدای آن مقیاس دارای مشکل بوده و نیاز به آموزش مجدد با تأکید روی سطوح پایین مقیاس دارد.

Tehran University Medical Journal; Vol. 64, No. 3, June 2006 : 5-11

Radiologists' intra-rater agreement and category distinguishability in diagnosis of ovarian mass by ultrasonography

A. Akbarzadeh Bagheban*[1]
G. Babaei [2]
A. Kazemnejad [2]
S. Faghihzadeh [2]
F. Baradaran Anaraki [3]
Z. Elahipanah [3]

1- Department of Biostatistics, Shaheed Beheshti University of Medical Sciences
2- Department of Biostatistics, Tarbiat Modarres University
3- Department of Radiology, Tehran University of Medical Sciences

* Department of Biostatistics, School of Paramedics, Ghods Squ., Darband Ave., Tehran , Iran, Po Box: 19395-4618, Tel: +98(21)22707347, Fax: +98(21)22721150
Email: akbarzad@sbum.ac.ir

ABSTRACT

Background: Intra-rater agreement in observing and decision making in diagnosis of any disease is of great importance.This investigation is to observe and read ultrasound pictures of ovarian cysts and distinguish its category for any radiologist. Distinguishability is one of the related entities in this matter and radiologists' ability in correct diagnosis is of great concern.
In this study, we evaluated radiologist's distinguishability of ordered categories of ovarian cyst diseases (benign, borderline and malignant) in ultrasonography. To do this, we measured intra-rater agreement of radiologists by Weighted Kappa coefficient, and then by the help of "square scores association model" and "agreement plus square scores association model" we evaluated their distinguishability in diagnosis of the severity of the ovarian cyst's diseases.
Methods: In this analytical cross-sectional study, two radiologists and three radiology residents assessed ultrasounds of 40 patients separately and independently in two periods (with the interval of one week). Patients selected from those who were referred to Mirza Koochak Khan Hospital in January 2005. Ultrasounds were performed by an expert radiologist and by a single apparatus.
Result: Data from radiologists was evaluated by "square scores association model" due to their superior results of distinguishability. Mean of Weighted Kappa coefficient was 0.81 and intra-rater agreement was 0.99 for our radiologists, but due to weaker results of our residents, we used "agreement plus square scores association model" for analyzing and mean of Weighted Kappa coefficient was 0.65 and intra-rater agreement was 0.97 for them.
Conclusion: Although radiologists had a better function than their residents, all of them showed appropriate distinguishability and intra-rater agreement in diagnosis and categorizing of the ovarian cyst's disease. To distinguish benign category from borderline was more difficult than to distinguish malignant category from borderline and radiologists showed better results in this than their residents did.

Keywords: Ovarian cyst, ultrasonography, reliability, weighted kappa, association model

References

1. Perkins SM, Becker MP. Assessing rater agreement using marginal association models. *Stat Med* 2002; 21: 1743-1760.
2. Koch GG, Landis JR, Freeman JL, Freeman DH, Lehnen RG. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 1977; 33: 133-158.
3. Agresti A. A model for agreement between ratings on an ordinal scale. *Biometrics* 1988; 44:539-548.
4. Kraemer HC, Periakoil VS, Noda A. Tutorial in biostatistics, kappa coefficients in medical research. *Stat Med* 2002; 21: 2109-2129.
5. Cohen J. A coefficient of agreement for nominal scales. *Education and Psychological Measures* 1960; 20: 37-46.
6. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; 70: 213-220.
7. Tanner MA, Young MA. Modelling ordinal scale agreement among raters. *JASA* 1985; 80: 175-180.
8. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; 43 :543-549.
9. Agresti A.Categorical Data Analysis. In: Agresti A.modles for matched paires.2 nd ed.New Jersey;2002: 409-42.
10. Goodman LA. Simple models for the analysis of association in cross-classifications having ordered categories. *JASA* 1979;74: 537-552.
11. May SM. Modelling observer agreement – an alternative to kappa. *J Clin Epidemiol* 1994; 44: 1315-24.
12. Fleiss JL. The design and Analysis of Clinical Experiment.In: Fleiss JL.Reliability of Meseaurment.1st ed. New York;1986: 1-28.
13. Darroch JN, McCloud PI. Category distinguishability and observer agreement. *Austr & New Zeal J of Statistics* 1986;28: 371-388.
14. Becker MP, Agresti A. Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Stat Med* 1992; 11: 101-114.
15. Byrt T. How good is that agreement ? (Letter to editor). *Epidemiology* 1996 ; 7 : 561.
16. Amer SA, Li TC, Bygrave C, Sprigg A, Saravelos H, Cooke ID. An evaluation of the inter-observer and intra-observer variability of the ultrasound diagnosis of polycystic ovaries. *Hum Reprod* 2002; 17: 1616-1622.
17. Stoker J, Desjardins P, Deleon A. Ultrasonography: its usefulness and reliability in early pregnancy: a review of 210 cases. *Am J Obstet Gynecol* 1975; 121: 1084-1088.